基于循环卷积神经网络的实体关系抽取方法研究 *

万 静[1]，李浩铭[1]，严欢春[1]，张雪超[2†]

(1. 北京化工大学 信息科学与技术学院，北京 100029; 2. 国防大学 联合勤务学院，北京 100091)

摘 要：针对目前大多数关系抽取中对于文本语料中较长的实体共现句，往往只能获取到局部的特征，并不能学习到长距离依赖信息的问题，提出了一种基于循环卷积神经网络与注意力机制的实体关系抽取模型。将擅长处理远距离依赖关系的循环神经网络 GRU 加入到卷积神经网络的向量表示阶段，通过双向 GRU 学习得到词语的上下文信息向量，在卷积神经网络的池化层采取分段最大池化方法，在获取实体对结构信息的同时，提取更细粒度的特征信息，同时在模型中加入基于句子级别的注意力机制。设计了在 NYT 数据集的实验验证，实验结果表明提出方法能有效提高实体关系抽取的准确率与召回率。

关键词：GRU；循环卷积神经网络；注意力机制；关系抽取

中图分类号：TP391 doi: 10.3969/j.issn.1001-3695.2018.09.0635

Relation extraction based on recurrent convolutional neural network

Wan Jing[1], Li Haoming[1], Yan Huanchun[1], Zhang Xuechao[2†]

(1. *College of Information Science & Technology, Beijing University of Chemical Technology, Beijing* 100029, *China*; 2. *College of Joint Logistics National Defence University, Beijing* 100091, *China*)

Abstract: Most of the relation extraction approaches could not learn the long distance dependence information from the long sentences with entity co-occurrence. This paper proposes a new relation extraction model to solve this problem. This model was based on the recurrent convolutional neural network and the sentence-level attention mechanism. It used the Bi-GRU neural network to learn context vectors for words. And it adopted the piecewise maximum pooling method, which could obtain fine grained features. This paper conducted experiments on the NYT dataset. Experimental results demonstrate that the proposed method outperforms the baseline systems.

Key words: GRU; recurrent convolutional neural networks; attention; relation extraction

0 引言

互联网的飞速发展，在带来海量数据的同时，也使得方便快捷提取出有效信息这一需求变得愈加重要；近年来，Freebase[1]、DBpedia[2]、YAGO[3]等知识库的建立，在 NLP 任务中得到广泛的应用。在该背景下，以从非结构化自然语言文本中，提取出结构化信息为目标的信息抽取技术应运而生。信息抽取的内容主要包括命名实体、关系和事件三类。实体关系抽取任务作为信息抽取的子任务之一，近些年来一直是学术界和工业界的研究热点，对于信息检索、自动问答、智能推荐等前沿领域都具有重要的意义。

传统的实体关系抽取方法需要人工设计特征、标注语料，耗费大量时间及人力，特征的选择直接影响到关系分类器的最终效果，且 NLP 标注工具的使用容易导致错误传播问题。近几年兴起的深度神经网络模型可以通过深层网络对大规模文本语料自动学习[4]，卷积神经网络因其优秀的特征提取能力已逐渐被用于实体关系抽取任务中[5]。然而，对于文本语料中较长的实体共现句，并不能学习到长距离依赖信息；并且对于实体关系抽取任务而言，普通最大池化方法虽然可以提取出最具价值的特征信息，但却无法捕获两个实体间的结构信息[6]。

针对目前实体关系抽取中简单卷积神经网络提取特征的局限性，本文主要研究基于循环卷积神经网络与注意力机制的实体关系抽取方法，通过结合循环神经网络和卷积神经网络，并加入注意力机制的方法来提高实体关系抽取的效果。

针对简单卷积神经网络不能学习长距离依赖信息的问题，本文提出将擅长处理远距离依赖关系的循环神经网络 GRU 加入到卷积神经网络的向量表示阶段。针对普通最大池化无法捕获两个实体间结构信息的问题，本文提出在卷积神经网络的池化层采取分段最大池化方法。在池化阶段，以两个实体为分隔点，将整个句子向量划分为三段，分别对每一段进行最大池化操作，之后再将三个池化向量拼接到一起。本文提出在关系抽取模型中加入基于句子级别的注意力机制，提高实体关系抽取的准确率。

本文将实体关系抽取任务看做多分类问题，研究结合使用循环神经网络和卷积神经网络的方法，以更准确地对实体间的关系进行分类；并通过增加注意力机制来提高实体关系抽取的效果。

1 基于 GRU_PCNN 和注意力机制的模型

本文提出的基于循环卷积神经网络和注意力机制的关系抽取方法包括三个阶段：基于双向 GRU 的向量表示阶段、基于 PCNN 的特征学习阶段和注意力权重学习阶段。

将原始输入句子转换为相应的词向量表示，将词向量与

收稿日期：2018-09-05；**修回日期：**2018-10-31 **基金项目：**国家科技支撑计划资助项目（2015BAK03B04）

作者简介：万静（1975-），女，湖北大悟人，副教授，博士，主要研究方向为自然语言处理，知识图谱；李浩铭（1994-），男，河南南阳人，硕士，主要研究方向为实体关系抽取、知识图谱；严欢春（1992-），女，山西运城人，硕士，主要研究方向为实体关系抽取、知识图谱；张雪超（1974-），男（通信作者），河南方城人，教授，博士，主要研究方向为大数据、知识图谱（zhangxuechaobj@163.com）.

词的位置特征拼接作为输入层的输入，句子输入后经过 Bi-GRU 层学习远距离依赖信息，将 GRU 的输出作为 PCNN 的输入，经过 PCNN 的特征学习后，通过句子级别的注意力机制来减少远程监督带来的噪声影响。最后，通过 softmax 进行归一化处理。本文提出的模型如图 1 所示。

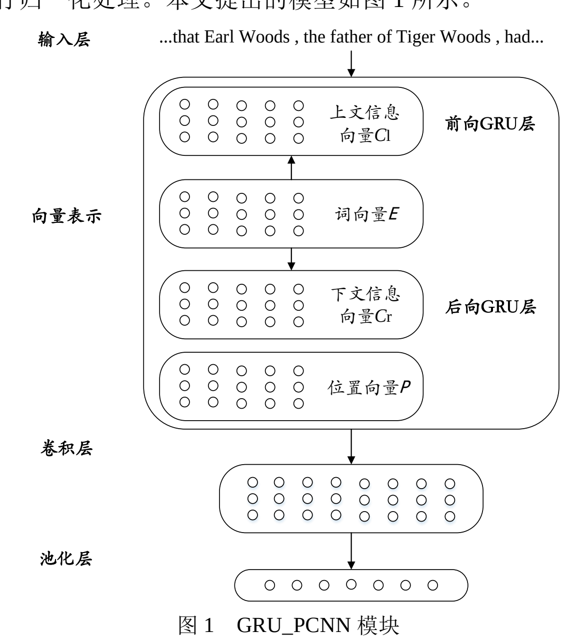


图 1　GRU_PCNN 模块

Fig. 1　GRU_PCNN module

1.1　基于 GRU 的向量表示

基于 GRU 的向量表示阶段主要负责将原始输入句子转换为神经网络需要的向量形式，以便进行后续的特征提取等一系列学习操作[7]。目前的实体关系抽取任务中，大多采用词语的分布表示——词向量作为神经网络的输入。本文在使用词向量的同时加入词语的位置向量，获取词语与实体间的相对位置信息为模型提供更多的特征。此外，本文提出将循环神经网络加入到向量表示阶段，借助循环神经网络的“记忆”能力来刻画上下文特征向量，将更丰富的特征提供给关系抽取模型。本文所说的“词语”不只包括单纯的一个单词，还包括实体词语，如“Tiger Woods”便是一个人名实体词语。

1.1.1 词分布表示

词分布表示（distributed representation）是基于分布假说中“词语的上下文分布决定词语的语义”这一思想，通过对词语的上下文建模来构造包含语义信息的词分布表示向量。词分布表示，也称为“词向量”“词嵌入”，其基本方法是通过模型的训练将文本中每个词语映射到一个新的空间，并以高密度、低维度的连续实数向量进行表示。对于一个包含 t 个词语的输入句子 $S=\{w_1,w_2,\ldots,w_t\}$，每个词语 w_i 将被转换为一个 d^w 维度的实数向量 $e(w_i)$。

1.1.2 位置特征表示

在关系抽取任务中，与两个实体距离比较近的词语，其对目标关系抽取的贡献就越大。因此，本文在向量表示阶段引入位置特征[8]，为实体关系抽取模型提供词语的位置结构信息，帮助模型判断目标关系词语的位置及类型。对于输入句子 S，通过计算当前词语 w_i 分别到两个实体 head entity、tail entity 的相对距离来得到其位置特征。以句子“…spread that Earl Woods , the father of Tiger Woods , had…”为例，实体词语“Earl Woods”和“Tiger Woods”分别是该句子的 head entity 和 tail entity，词语“father”到这两个实体的距离分别为 3、-2。在本文的模型中，词语 w_i 到两个实体的相对距离会被映射转换成 d^p 维度的向量 d_1、d_2，然后组合得到词语的位置特征向量 $P(w_i)=[d_1,d_2]$。

1.1.3 基于 Bi-GRU 的上下文信息向量表示

循环神经网络（RNN）因其擅长刻画文本序列的特性被用于许多自然语言处理任务中，并取得了不错的研究效果。然而，梯度消失使得 RNN 很难学习到句子中的远距离依赖信息。于是，擅长学习远距离依赖信息的长短期记忆网络 LSTM 被提出[10]。LSTM 通过三个门的操作对已学习到的序列信息实现记忆与忘记功能。其中，遗忘门负责决定上一时刻学习到的信息保留多少到当前时刻；输入门负责决定当前时刻的输入保留多少用于当前时刻的学习；输出门则负责决定输出多少当前时刻学习到的信息。GRU（gated recurrent unit）是 Cho 等人[11]2014 年提出的循环神经网络模型，是对 LSTM 复杂网络结构进行简化的一种变体。GRU 模型将 LSTM 模型中遗忘门和输入门合并成单一的更新门，研究表明 GRU 在模型参数减少的情况下可以获得比 LSTM 更好的效果[14]。因此，本文使用 GRU 模型来学习词语的上下文信息。

GRU 模型中包含两个门：重置门和更新门，即图 2 中的 r_i 和 z_i。对于句子中的词语 w_i，学习其上文信息向量 $c_l(w_i)$ 时，需要用到上一个词语的分布表示向量和上文信息向量 $e(w_{i-1})$、$c_l(w_{i-1})$，则模型各部分的计算公式分别为

$$r_i^{(l)}=\sigma(W_r^{(l)}\cdot[c_l(w_{i-1});e(w_{i-1})]) \tag{1}$$

$$z_i^{(l)}=\sigma(W_z^{(l)}\cdot[c_l(w_{i-1});e(w_{i-1})]) \tag{2}$$

$$c_l(w_i)=\tanh\left(W_{c_l}\cdot[r_i^{(l)}\cdot c_l(w_{i-1});e(w_{i-1})]\right) \tag{3}$$

$$c_l(w_i)=(1-z_i^{(l)})*c_l(w_{i-1})+z_i^{(l)}*c_l(w_i) \tag{4}$$

其中：$[c_l(w_{i-1});e(w_{i-1})]$ 是对向量 $c_l(w_{i-1})$ 和 $e(w_{i-1})$ 进行拼接。计算句子第一个词语的上文信息向量时需要用到 $c_l(w_{-1})$ 和 $e(w_{-1})$，本文在初始化时将它们都设定为零向量。$W_r^{(l)}$、$W_z^{(l)}$、W_{c_l} 分别为三个公式的权重矩阵，因参与计算的向量是拼接而成的，所以权重矩阵在学习时也是分开的，即

$$W_r^{(l)}=W_{rc}^{(l)}+W_{re}^{(l)} \tag{5}$$

$$W_z^{(l)}=W_{zc}^{(l)}+W_{ze}^{(l)} \tag{6}$$

$$W_{c_l}=W_{c_lc}+W_{c_le} \tag{7}$$

其中：σ 表示 sigmoid 函数，其结果介于 0~1。对于重置门，当 sigmoid 函数结果非常接近 0 时，表示上文信息将被忽视，当前输入信息被重置。这样的设定可以帮助人们舍弃无关信息，得到更简洁、更有价值的向量表示。更新门的主要作用是控制参与当前学习的上文信息数量，它可以帮助模型记忆长距离依赖信息。

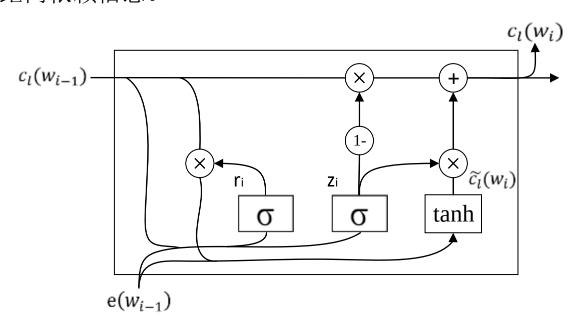


图 2　前向传播 GRU 模型

Fig. 2　Forward propagation GRU model

图 2 所展示的前向传播 GRU 模型可以帮助人们记忆上文信息，以便学习前文中的远距离依赖关系。本文设计的向

量表示学习模块不仅要对词语的上文信息进行记忆学习，还要学习下文的远距离依赖信息。因此本文所用的循环神经网络是双向 GRU（Bi-GRU）模型，前向 GRU 用于词语 w_i 的上文信息向量 $c_l(w_i)$，后向 GRU 用于学习下文信息向量 $c_r(w_i)$。以句子“…spread that Earl Woods , the father of Tiger Woods , had…”为例，词语“father”的上文信息向量 $c_l(w_i)$ 是对其左边文本“…spread that Earl Woods , the”学习记忆的结果，下文信息向量 $c_r(w_i)$ 是对该词语右边文本“of Tiger Woods , had…”学习记忆的结果。

下文信息向量的学习与前文类似，只是其输入是逆序的，故计算时先对最后一个词语进行下文信息学习，再依次往前，计算公式如下：

$$r_i^{(r)}=\sigma\left(W_r^{(r)}\cdot[c_r(w_{i+1});\mathrm{e}(w_{i+1})]\right) \tag{8}$$

$$z_i^{(r)}=\sigma\left(W_z^{(r)}\cdot[c_r(w_{i+1});\mathrm{e}(w_{i+1})]\right) \tag{9}$$

$$c_l(w_i)=\tanh\left(W_{c_r}\cdot[r_i^{(r)}\cdot c_r(w_{i+1});\mathrm{e}(w_{i+1})]\right) \tag{10}$$

$$c_r(w_i)=(1-z_i^{(r)})*c_r(w_{i+1})+z_i^{(r)}*c_r(w_i) \tag{11}$$

相应地，权重矩阵 $W_r^{(r)}$、$W_z^{(r)}$、W_{c_r} 在学习过程中也是分开的，即

$$W_r^{(r)}=W_{rc}^{(r)}+W_{re}^{(r)} \tag{12}$$

$$W_z^{(r)}=W_{zc}^{(r)}+W_{ze}^{(r)} \tag{13}$$

$$W_{c_r}=W_{c_rc}+W_{c_re} \tag{14}$$

经过这一系列计算，本文得到了词语 w_i 的词向量 $\mathrm{e}(w_i)$、位置向量 $\mathrm{P}(w_i)$、上文信息向量 $c_l(w_i)$、下文信息向量 $c_r(w_i)$，将它们进行拼接，最终得到包含语义特征、上下文特征及位置特征的向量表示：$x_i=[c_l(w_i);\mathrm{e}(w_i);c_r(w_i);\mathrm{P}(w_i)]\in R^d$。其中，$d=d^l+d^w+d^r+2*d^p$，$d^l$、$d^r$ 分别为上下文信息向量的维度。

1.2　基于 PCNN 的特征学习

1.2.1 卷积

在关系抽取任务中，通过对两个目标实体共同出现的句子进行学习，以预测它们之间的关系。为此，需要对整个句子进行特征抽取。本文采用擅长提取特征的卷积神经网络来对向量表示层获取到的所有特征进行学习，以更好的进行关系预测。

卷积层通过对滑动窗口内的向量进行卷积操作达到特征提取的效果。对于句子 $\mathrm{s}=\{x_1,x_2,\ldots,x_t\}$（其中 x_i 是上一阶段学习到的词语 w_i 的向量表示），定义 $x_{i:j}$ 表示序列 $[x_i:x_{i+1}:\ldots:x_j]$ 的拼接，l 为卷积核滑动窗口的长度，则有卷积矩阵 $W_i\in R^{l\times d}$。

通常为了学习到多种特征，模型中会使用多个卷积核。假设本文的模型中使用了 n 个卷积核，则卷积矩阵为 $W=\{W_1,W_2,\ldots,W_n\}$，卷积操作可以表示如下：

$$c_{ij}=W_i\otimes x_{j-l+1:j}\ (1\le \mathrm{i}\le \mathrm{n},1\le j\le t+l-1) \tag{15}$$

考虑到当 j 接近 1 或 t 时，滑动窗口可能会超出句子的边界，本文将所有超出句子范围的向量 $x_j\,(j<1\ \text{or}\ j>m)$ 均设定为零向量。

经过卷积操作后，可以得到卷积层的学习结果向量 $\mathrm{C}=\{c_1,c_2,\ldots,c_n\}$。

1.2.2 分段最大池化策略

池化层设置在卷积层之后，对卷积层学习到的特征进行进一步提取，在降低神经网络复杂度的同时，提取出主要特征。一般采用的池化方法是最大池化策略，该策略在卷积层每个卷积核学习到的一系列特征中选取最大值，提取出最有价值的特征，已被用于各类自然语言处理任务中。然而对于实体关系抽取来说，最大池化策略不能捕获两个实体间的结构信息，并且无法提取到细粒度的特征信息。因此，本文采用分段最大池化方法[9]，以两个目标实体为分隔点将整个句子划分为三段，分别对每一段进行最大池化操作。相应地，每一个卷积核的结果向量 c_i 会被分为三段 $\{c_{i1},c_{i2},c_{i3}\}$，分段最大池化的操作可以表示为

$$p_{ij}=\max(c_{ij})\quad 1\le i\le n,1\le j\le 3 \tag{16}$$

之后将三个池化向量结合到一起得到向量 $p_i=\{p_{i1},p_{i2},p_{i3}\}$。连接所有的 p_i 并进行非线性函数运算，便可得到池化层的最终输出向量为

$$\mathrm{s}=\tanh(p_{1:n}) \tag{17}$$

其中 $s\in R^{3n}$。可以看出句子的特征向量在经过池化操作后，已经成为一个与原句子长度无关的固定长度向量。

1.3　基于句子级别的注意力机制

1.3.1 注意力机制层

基于句子级别的注意力机制[15]（attention）层，会为每个实体共现句学习注意力值，即权重。正确表达目标关系的句子将获得较高的权重，而那些被错误标注的句子会得到非常低的权重。

对于一组实体对（e1，e2），所有它们共同出现的句子组成集合 $\mathrm{T}=\{s_1,s_2,\ldots,s_q\}$，Attention 层会为该集合计算相应的权重向量 $\alpha=\{\alpha_1,\alpha_2,\ldots,\alpha_q\}$。于是，集合 T 的特征便可计算如下：

$$\bar{T}=\sum_{i=1}^{q}\alpha_i s_i \tag{18}$$

本文通过计算句子特征表示向量与目标关系的相似度来得到句子的权重[16]。Bordes 等人[17]2013 年提出 TransE 的方法用于表示知识图谱中的实体关系，并已在许多知识图谱任务中取得了不错的效果。TransE 方法将知识图谱中的关系定义为头实体 e_1 到尾实体 e_2 的映射转换 $e_1+r\approx e_2$。基于这样的思想，当给定一个头实体 entity1 时，可以根据知识图谱中的关系表示预测出相应的尾实体 entity 2；同样地，也可根据关系表示和尾实体来预测出头实体。本文借鉴 TransE 的思想来表示实体间的关系：$r=e_1-e_2$，同时认为当实体共现句的特征表示向量 s_i 与目标关系向量 r 相似度越高时，该句子能正确表达目标关系的可能性越大，其注意力权重 α_i 也越高[12]。本文定义权重的计算如下：

$$\alpha_i=\frac{\exp(\omega_i)}{\sum_{j=1}^{q}\exp(\omega_j)} \tag{19}$$

其中：ω_i 是句子特征向量 s_i 与预测关系向量 r 的匹配分数，通过式(20）计算求得。

$$\omega_i=W_a(\tanh[s_i;r])+b_a \tag{20}$$

其中：$[s_i;r]$ 表示两个向量的垂直连接，W_a 为中间矩阵，b_a 是偏置向量。

1.3.2 Softmax 层

Softmax 是一种多分类模型，它是逻辑回归模型在多分类问题上的推广。在关系多分类问题中，关系类别标签会对应多个不同的值。添加了注意力权重的句子特征表示被送入 Softmax 层，计算其对应所有关系类型的匹配分数：

$$o=W_s\bar{T}+b_s \tag{21}$$

其中：$o\in R^{n_o}$ 表示模型的输出，$W_s\in R^{n_o\times 3n}$ 为权重矩阵，$b_s\in R^{n_o}$ 是偏置向量。本文定义第 i 个关系类型的条件概率为

$$p(r_i|T;\theta)=\frac{\exp(o_i)}{\sum_{j=1}^{n_o}\exp(o_j)} \tag{22}$$

其中：T 表示句子向量的集合，n_o 为关系总数量，θ 代表模型中的所有参数的集合。

假设训练数据中共有 m 个句子集合，每个集合代表一类关系。于是，可以定义目标函数如下：

$$J(\theta)=\sum_{i=1}^{m}logp(r_i|T_i;\theta) \tag{23}$$

2　实验及结果分析

2.1　实验设计

2.1.1 实验数据

关系抽取研究中常用的知识图谱是 Freebase，它是一个包含 4 000 多万实体、上万个属性关系、24 多亿个事实三元组的大规模知识图谱。相应地，文本语料选用 New York Times（NYT）数据集，该数据集涵盖了 1987—2007 年期间所有的纽约时报新闻，并且包含了大量 Freebase 中的实体，通过远程监督得到数据集，可以为人们提供丰富的实体共现句。Freebase 和 New York Times 数据集非常庞大，模型的训练测试并不需要使用所有的数据。本文使用 2010 年 Riedel 等人将 NYT 语料与 Freebase 对齐生成的实体关系标注数据集来进行模型的实验。

标注好的 NYT 语料分为两部分，一部分是 2005—2006 年间的新闻语料，包含 281 270 组实体对和 522 611 个实体共现句，作为模型的训练数据；另一部分是 2007 年期间的新闻语料，包含 96 678 组实体对和 172 448 个实体共现句，用于模型的测试。语料中的关系类型共 53 种，如图 3 所示，其中“NA”表示两个实体间不存在关系。表 1 展示了数据集中的关系实例。

/business/shopping_center_owner/shopping_centers_owned	/location/province/capital
/location/neighborhood/neighborhood_of	/people/person/nationality
/location/fr_region/capital	/business/person/company
/location/cn_province/capital	/location/mx_state/capital
/location/in_state/administrative_capital	/business/company/advisors
/base/locations/countries/states_provinces_within	/business/shopping_center/owner
/business/company/founders	/people/person/ethnicity
/location/country/languages_spoken	/people/deceased_person/place_of_burial
/people/person/place_of_birth	/people/ethnicity/geographic_distribution
/people/deceased_person/place_of_death	/people/person/place_lived
/location/it_region/capital	/business/company/major_shareholders
/people/family/members	/broadcast/producer/location
/location/us_state/capital	/broadcast/content/location
/location/us_county/county_seat	/business/business_location/parent_company
/people/profession/people_with_this_profession	/location/jp_prefecture/capital
/location/br_state/capital	/film/film/featured_film_locations
/location/in_state/legislative_capital	/people/place_of_interment/interred_here
/sports/sports_team/location	/location/de_state/capital
/people/person/religion	/people/person/profession
/location/in_state/judicial_capital	/business/company/locations
/business/company_advisor/companies_advised	/location/country/capital
/people/family/country	/location/location/contains
/time/event/locations	/location/country/administrative_divisions
/business/company/place_founded	/people/person/children
/location/administrative_division/country	/film/film_location/featured_in_films
/people/ethnicity/included_in_group	/film/film_festival/location
NA	

图 3　数据集中的关系类型

Fig. 3　Relation types in dataset

表 1　数据集中的关系实例

Table 1　Instances in dataset

实体 Aid/ 实体 Bid	实体 A/ 实体 B	关系	句子
m.0ccvx m.05gf08	queens belle_harbor	/location/location /contains	…into the fatal crash of a passenger jet in belle_harbor , queens , …
m.05kkh m.0y_kj	Ohio celin	/location/location /contains	where : celina , ohio .

2.1.2 评价指标

实体关系抽取任务中通常采用准确率 P、召回率 R 和 *F*-measure（*F*1）作为评价指标，它们的计算公式如下：

$$P=\frac{正确抽取的实体关系数量}{抽取的实体关系总数量} \tag{24}$$

$$R=\frac{正确抽取的实体关系数量}{样本中的实体关系总数量} \tag{25}$$

$$F1=\frac{2*P*R}{P+R} \tag{26}$$

在此基础上，本文也采用了准确率—召回率曲线（PRC）以更直观地展示与其他算法的对比结果。

2.1.3 词向量

本文的实体关系抽取模型中，词向量不仅是提取特征的 PCNN 输入向量的重要部分，更是学习上下文信息的循环神经网络 GRU 的输入。若使用随机初始化的词向量，模型训练所产生的效果可能不稳定，不如通过词分布表示模型在大规模语料上训练的词向量。Word2vec[13]是 Google 公司于 2013 年开源的词向量训练工具，它实现了 skip-gram 和 CROW 模型。其中 skip-gram 模型在表达词语语义关系方面效果更优一些。因此，本文采用 word2vec 工具和 Skip-gram 模型在 New York Times 文本数据集上进行英文词向量的训练学习。

2.1.4 实验参数

本文使用交叉验证的方法在训练集上对模型进行调优。模型中各参数的范围分别设定为：词分布表示向量的维度在 $\{50,100,200,300\}$ 中取值，位置特征向量的维度在 $\{5,10,20\}$ 中选择，卷积滑动窗口大小的取值为 $\{3,5,7\}$，卷积核的数量选择为 $\{50,100,150,200,250\}$，小批量梯度下降的学习速率取值为 $\{0.1,0.01,0.001\}$，每个批次的数量取值为 $\{50,100,150,200\}$。经过实验，模型最终的参数设置为如表 2 所示。

表 2　实验参数设置

Table 2　Experimental parameter settings

参数	参数值
词向量维度	50
上下文信息向量维度	50
位置特征向量维度	5
卷积滑动窗口大小	3
卷积核数量	200
批次数量	50
学习速率	0.01

2.1.5 加入双向 GRU 的影响

本文提出了将擅长学习远距离序列依赖信息且网络结构简单的循环神经网络 Bi-GRU 加入到向量表示阶段，分别通过前向、后向 GRU 学习词语的上文信息向量和下文信息向量，为后续的卷积学习提供了丰富的特征信息。为了验证本文方法的有效性，设计一组基于分段卷积神经网络（PCNN_ATT）与加入双向 GRU（GRU_PCNN_ATT）的实体关系抽取对比实验。为了方便实验结果的比较，本文为单独的 PCNN 模型也加入基于句子级别的注意力机制。此外，PCNN_ATT 模型的向量输入由词向量和位置特征向量组成，而 GRU_PCNN_ATT 模型的向量输入不仅包括词向量和位置特征向量，还包括经由双向 GRU 获得的上下文信息向量。

2.1.6 加入句子级别注意力机制的影响

本文提出了加入基于句子级别的 Attention 机制，对每个句子计算其注意力权重，增大正确表达目标关系语句的权重，同时减小错误标注语句的权重，降低其对模型训练的干扰。

本文设计实验对比 GRU_PCNN_ONE、GRU_PCNN_AVG、GRU_PCNN_ATT 三种情况下的实体关系抽取效果。其中，GRU_PCNN_ONE 是指对于一组实体对，在它们的共现句集合中随机选择一个句子进行关系的学习；GRU_PCNN_AVG 是指赋予一组实体对的所有共现句相同的权重去参与关系的学习；GRU_PCNN_ATT 是本文提出的基于句子级别注意力机制的实体关系抽取模型，对每一个实体共现语句分别计算注意力权重，每个句子以相应的权重去参与关系的学习。

2.2 实验结果分析

2.2.1 双向 GRU 影响分析

在表 3 和图 4 中可以看到，加入双向 GRU 的关系抽取模型比普通的分段卷积神经网络模型的效果更优，在整个召回率范围内，GRU_PCNN_ATT 方法的准确率比基于分段卷积神经网络(PCNN_ATT)方法大约要高出 0.05。这是因为 GRU 作为一种循环神经网络，具备优秀的记忆序列特征的能力；同时，它又像 LSTM 一样，善于学习长距离依赖信息。实验语料中存在许多长依赖语句，仅靠滑动窗口获得局部上下文信息的 PCNN_ATT 模型效果不能捕捉到长依赖信息，然而拥有记忆能力的双向 GRU_PCNN_ATT 模型却可以学习到丰富的上下文特征，其结果便更优一些。

表 3 加入 Bi-GRU 的实验对比结果

Table 3 Experimental results of adding Bi-GRU

方法	准确率（%）	召回率（%）	F1 值（%）
CNN_ATT	70.23	47.51	56.68
GRU_PCNN_ATT	72.5	50.08	59.24

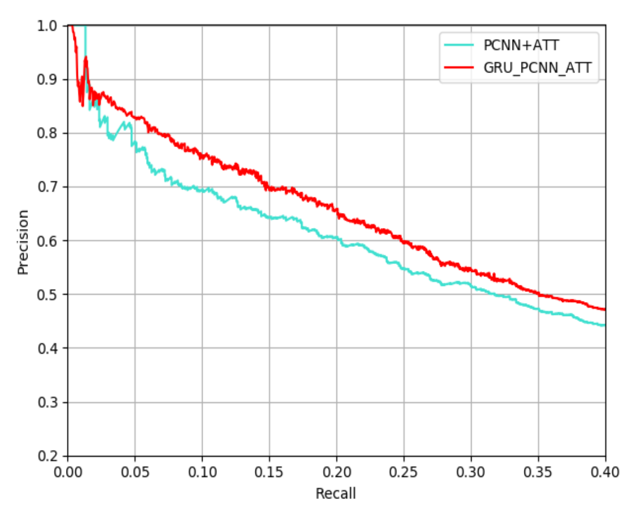


图 4 加入 Bi-GRU 的实验对比结果

Fig. 4 Experimental results of adding Bi-GRU

2.2.2 注意力机制的影响分析

基于句子级别注意力机制的实验结果如表 4 所示，表中显示的数据为准确率值，单位均为“%”。可以看出，不论测试集大小如何，本文加入注意力机制的方法（GRU_PCNN_ATT）效果均高出随机选择（GRU_PCNN_ONE）和平均权重（GRU_PCNN_AVG）约 3%-4%，说明基于句子级别注意力机制的加入有助于实体关系抽取准确率的提升。此外，由表中数据可以看出使用平均权重的模型（GRU_PCNN_AVG）准确率要低许多，尤其是当测试集为全部语句时。相对于随机选择的模型（GRU_PCNN_ONE）而言，使用平均权重的模型虽然学习的语句更多，可以获得更多的特征，但同时也给予那些错误标注语句相同的权重值，导致噪声数据的混入，影响了模型的效果。

表 4 注意力机制实验结果

Table 4 Experimental results of adding Attention

测试集	ONE/%	THREE/%	ALL/%
GRU_PCNN_ONE	65.8	67.6	69.1
GRU_PCNN_AVG	65.3	67.5	68.2
GRU_PCNN_ATT	68.1	71.9	72.5

2.2.3 GRU_PCNN_ATT 方法验证及结果分析

本文的实体关系抽取方法与 Mintz、MultiR、MIML 这三种经典的远程监督方法的实验对比结果如表 5 和图 5 所示。表 3~5 展示了中文语料下各方法的实验结果，可以看出，不论是准确率还是召回率，本文方法皆远高于三种传统远程监督方法。图 5 中展示了英文语料每种方法的 PRC 曲线，可以看出在整个召回率范围内，本文方法（GRU_PCNN_ATT）的准确率均比三种基于特征的传统远程监督方法高出很多。当召回率大于 0.1 时，基于特征的关系抽取方法准确率大幅下降，而本文的方法还能取得不错的准确率。这是因为人工设计的特征不能够准确把握句子的语义信息，而且标注特征的 NLP 工具不可避免的会产生错误影响到关系抽取的效果。相比来说，本文基于神经网络的关系抽取方法可以避免一些 NLP 工具的错误，更加准确地学习到句子的语义信息。

表 5 与传统远程监督方法对比结果

Table 5 Comparison with traditional distant supervision methods

方法	准确率（%）	召回率（%）	F1 值（%）
Mintz	62.35	36.82	46.30
MultiR	65.43	41.31	50.64
MIML	66.52	43.70	52.75
GRU_PCNN_ATT	72.5	50.08	59.24

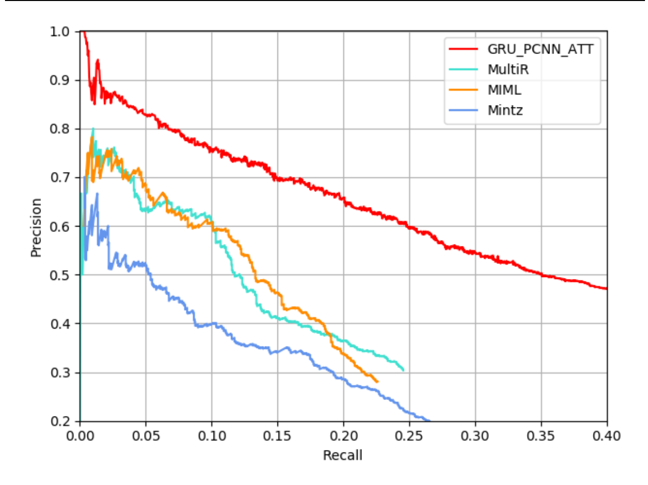


图 5 与传统远程监督方法对比结果

Fig. 5 Comparison with traditional distant supervision methods

3 结束语

本文基于对现有关系抽取方法上所存在缺陷的分析提出了基于循环卷积神经网络和注意力机制的实体关系抽取方法，该方法通过双向 GRU 模型来学习词语的上下文信息，使用循环卷积神经网络获取更细粒度的特征信息，并提取到实体间的结构信息，同时加入基于句子级别的注意力机制。

通过一系列实验对比，证明了本文方法能有效提升实体关系抽取的效果。

参考文献：

[1] Huang Xun, You Hongliang, Yu Yang. A review of relation extraction [J]. New Technology of Library & Information Service, 2013(11):

30-39

[2] Kurt B, Colin E, Praveen P, *et al*. Freebase: a collaboratively created graph database for structuring human knowledge [C]// Proc of KDD. 2008: 1247-1250.

[3] Lehmann J. DBpedia: a nucleus for a web of open data [C]// Proc of Semantic Web, International Semantic Web Conference&Asian Semantic Web Conference. 2007: 11-15.

[4] Liu Chunyang, Sun Wenbo, Chao Wenhan, *et al*. Convolution neural network for relation extraction [C]//Proc of International Conference on Advanced Data Mining and Applications. Berlin:Springer,2013: 231-242.

[5] Lin Yankai, Shen Shiqi, Liu Zhiyuan, *et al.* Neural relation extraction with selective attention over instances [C]//Proc of the 54th Annual Meeting of the Association for Computational Linguistics,Volume 1: Long Papers. 2016:2124-2133.

[6] Zeng Daojian, Liu Kang, Lai Siwei, *et al.* Relation classification via convolutional deep neural network [C]//Proc of the 25th International Conference on Computational Linguistics:Technical Papers.2014: 2335-2344.

[7] Zhou Peng, Shi Wei, Tian Jun, *et al*. Attention-based bidirectional long short-term memory networks for relation classification [C]// Proc of the 54th Annual Meeting of the Association for Computational Linguistics,Volume 2: Short Papers.2016:207-212.

[8] Nguyen T H, Grishman R. Relation extraction: perspective from convolutional neural networks [C]//Proc of the 1st Workshop on Vector Space Modeling for Natural Language Processing. 2015: 39-48.

[9] Zeng Daojian, Liu Kang, Chen Yubo, *et al*. Distant supervision for relation extraction via piecewise convolutional neural networks [C]// Empirical Methods in Natural Language Processing. 2015: 1753-1762.

[10] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.

[11] Cho K, Van Merrienboer B, Gulcehre C, *et al.* Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]//Proc of Empirical Methods in Natural Language Processing. 2014:1724-1734.

[12] Shen Yatian, Huang Xuanjing. Attention-based convolutional neural network for semantic relation extraction[C]// Proc of the 26th International Conference on Computational Linguistics: Technical Papers. 2016: 2526-2536.

[13] Mikolov T, Sutskever I, Chen Kai, *et al.* Distributed representations of words and phrases and their compositionality [C]//Advances in Neural Information Processing Systems.2013: 3111-3119.

[14] Jozefowicz R, Zaremba W, Sutskever I. An empirical exploration of recurrent network architectures[C]// Proc of International Conference on Machine Learning. 2015: 2342-2350.

[15] Mnih V, Heess N, Graves A. Recurrent models of visual attention [C]// Advances in Neural Information Processing Systems. 2014: 2204-2212.

[16] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [C]// Proc of ICLR. 2015.

[17] Bordes A, Usunier N, Garcia-Duran A, *et al.* Translating embeddings for modeling multi-relational data [C]// Neural Information Processing Systems. 2013: 2787-2795.